Independent Component Analysis for the identification of sources of variation on an industrial NIRS application

A. Moya-González[1], P. Barreiro[1], D. Jouan-Rimbaud-Bouveresse[2,3], D.N. Rutledge[2,3]

[1] *LPF-Tagralia, Universidad Politécnica de Madrid, Madrid, Spain, adolfo.moya@upm.es*

[2] *INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005, Paris*

[3] *AgroParisTech, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris*

Keywords: Industrial application, robustness, interference variation.

1- Introduction

A Near Infrared Spectroscopy (NIRS) industrial application was developed by the LPF-Tagralia team, and transferred to a Spanish dehydrator company (Agrotécnica Extremeña S.L.) for the classification of dehydrator onion bulbs for breeding purposes. The automated operation of the system has allowed the classification of more than one million onion bulbs during seasons 2004 to 2008 (Table 1).

Season	Classified onion bulbs
2004	284.964
2005	203.426
2006	114.550
2007	321.650
2008	281.202
Total	1.205.792

Table 1: *Number of bulbs classified by the automated system (2004-2008). During this period, a SSC increase of 0.241 ºBrix per season has been achieved (based in the onions measured in the framework of the breeding program).*

The performance achieved by the original model (R2=0,65; SEC=2,28ºBrix) was enough for qualitative classification thanks to the broad range of variation of the initial population (18ºBrix). Nevertheless, a reduction of the classification performance of the model has been observed with the passing of seasons. One of the reasons put forward is the reduction of the range of variation that naturally occurs during a breeding process, the other is the variations in other parameters than the variable of interest but whose effects would probably be affecting the measurements [1].

This study points to the application of Independent Component Analysis (ICA) on this highly variable dataset coming from a NIRS industrial application for the identification of the different sources of variation present through seasons.

2- Material and methods

The automated system estimates the onions Soluble Solids Content (SSC) by means of a NIR spectrometer (Hamamatsu PMA-11) equipped with an InGaAs sensor that measures reflectance at 244 channels, from 894 nm to 1649 nm. The spectrometer is PC controlled and the estimation model embedded is a MLR model calibrated from at-line measurements acquired in 2002. A later study [2] demonstrated that the wavelengths used by the MLR model were also selected when calibrating a PLS model with new data included.

The spectra were preprocessed with first order derivative using the Savitzky–Golay smoothing filter, a wavelength pruning of the right tail of the spectra (over 1427 nm) that was demonstrated as a noisy region with no informative value [3], and finally with an SNV transformation.

ICA calculations were done using the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [4] downloaded from http://perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m, and in-house codes for the validation methods. All computations were performed using Matlab, The MathWorks Inc., Natick (MA, USA).

3- Results and discussion

The study of the correlations between the extracted ICs has been established by Jouan-Rimbaud--Bouveresse *et al*. in a study not yet published. The so-called *ICA-By-Blocks* procedure provides Signal-Correlation graphs that could be used in the determination of the total number of true ICs present on a data set. The procedure divides the data to be analyzed into a number of blocks, which are representative of the whole dataset, and calculates ICA models for each block, from 1 to a selected maximum number of ICs. If the extracted ICs corresponds to 'true ICs', they should be present in the different blocks and highly correlated with the equivalent ones extracted for the other blocks. The *ICA-By-Blocks* procedure was performed for the whole dataset, the Signal-Correlation plot (Figure 1) shows that only two ICs are consistently extracted from the three blocks. When plotting the signals corresponding to the three different blocks for the optimal number of ICs to be extracted, it can be seen that the extracted signals for each IC are very similar, as they were supposed to be equivalent (see Figure 2 and Figure 3). When extracting 2 ICs from the 2008 dataset the signals were equivalent for the three blocks (see Figure 4 and Figure 5). In all cases, the blocks were defined by a Venetian Blind algorithm.

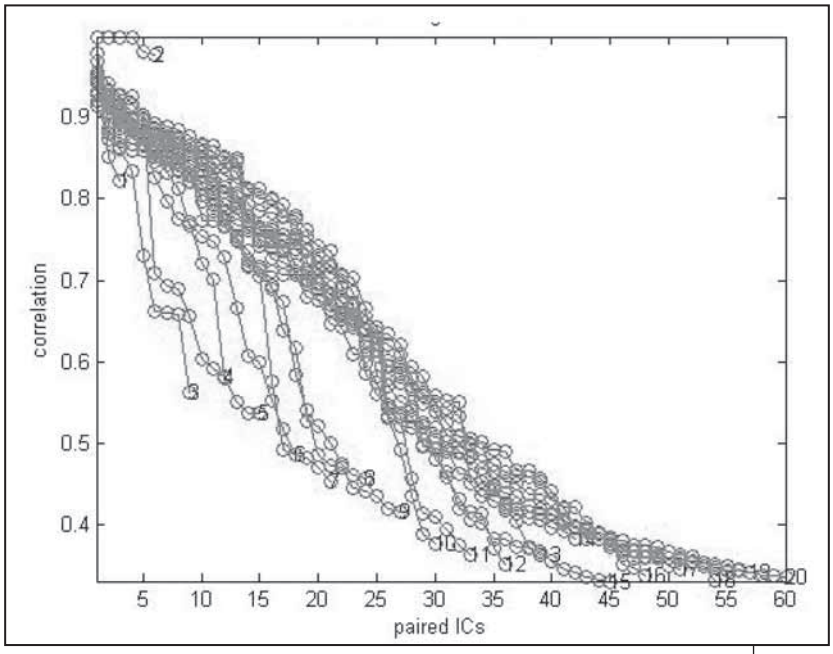

Figure 1: *Signal-Correlation graph for the whole dataset (3 Blocks, 1 to 20 ICs). Only for 2 ICs the correlation coefficient for the paired signals is over 0.95. $n \times (B^2 - B)$ correlation coefficients are plotted for each IC model, being n the number of ICs extracted and B the number of blocks.*

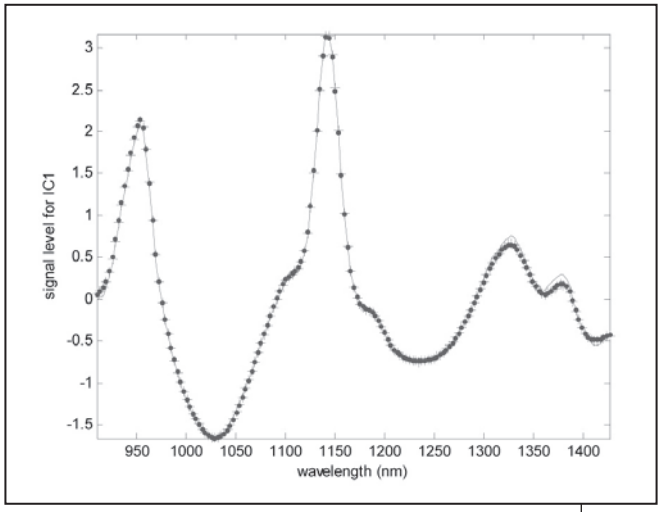

Figure 2: *Signals for IC1 extracted from the three different blocs. Block 1 (solid line), block 2 (pluses) and block 3 (dotted line). 2 ICs extracted from the whole dataset (1.205.792 samples)*

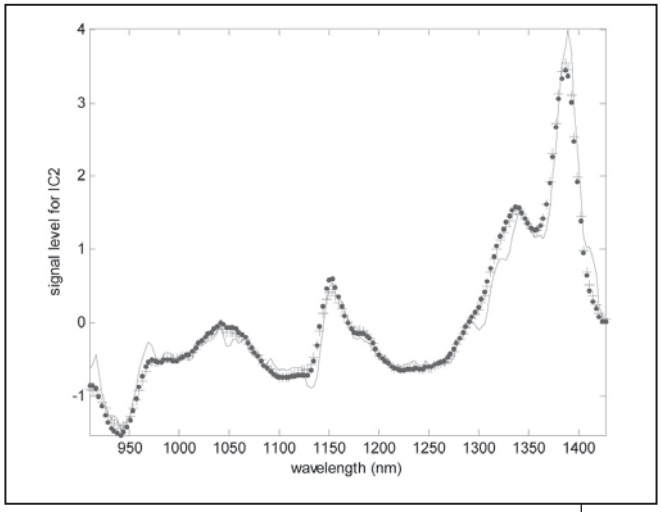

Figure 3: *Signals for IC2 extracted from the three different blocs. Block 1 (solid line), block 2 (pluses) and block 3 (dotted line). 2 ICs extracted from the whole dataset (1.205.792 samples)*

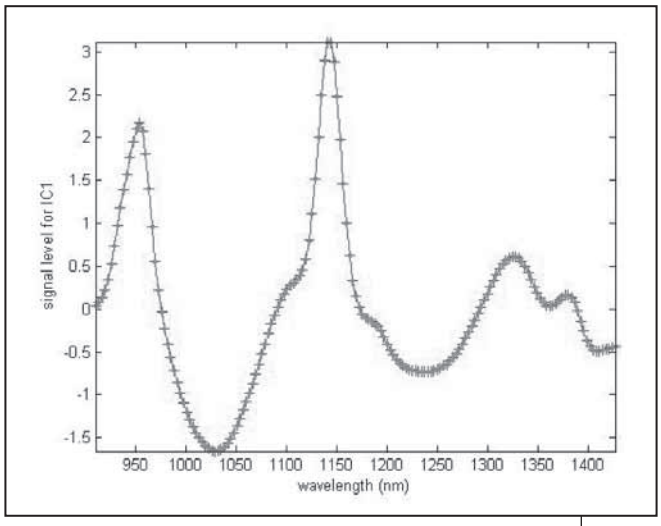

Figure 4: *Signals for IC1 extracted from the three different blocs. Block 1 (solid line), block 2 (pluses) and block 3 (dotted line). 2 ICs extracted from the 2008 dataset (281.202 samples)*

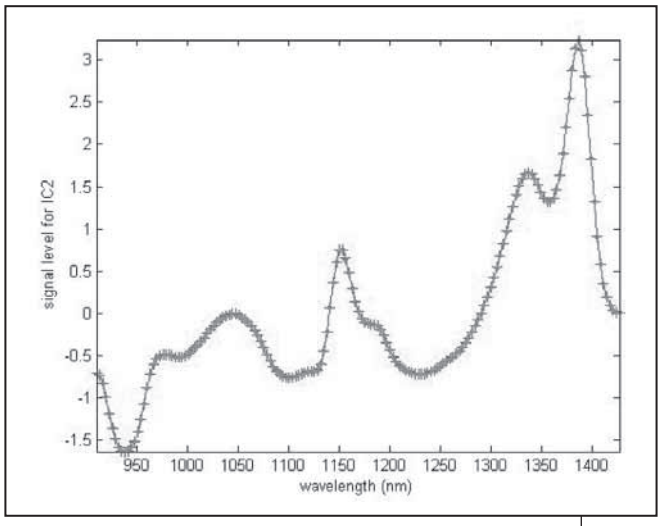

Figure 5: *Signals for IC1 extracted from the three different blocs. Block 1 (solid line), block 2 (pluses) and block 3 (dotted line). 2 ICs extracted from the 2008 dataset (281.202 samples)*

As many different sources of variation, including SSC, texture, light intensity, temperature, season, orchard etc., are known to be present and probably affecting the NIRS measurements, two ICs seems to be very few. When applying the *ICA-By-Blocks* procedure to the data by seasons, with the same number of blocks and maximum number of ICs extracted, the results show that, for each season, 8 to 10 significant ICs could be extracted. As an example, the correlations between signals have been plotted for the 2008 season where, based on the correlations of the paired ICs, the optimal number of ICs to be extracted is 8 or 9 (Figure 6).

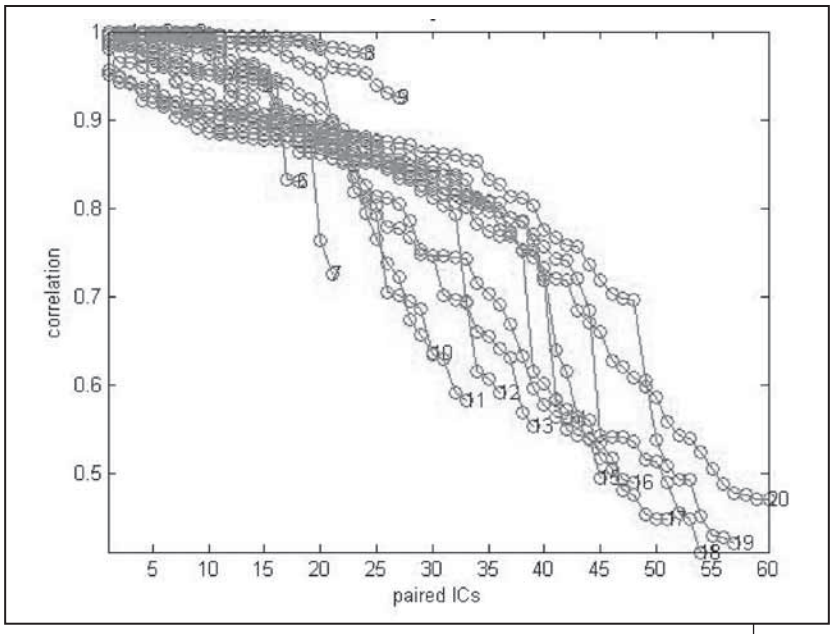

Figure 6: *Signal-Correlation graph for the 2008 data (3 Blocks, 1 to 20 ICs). For 8 ICs the correlation coefficient remains over 0.95, decreasing for a bigger number of ICs.*

The scores of 76 spectra acquired during the 2008 season with SSC reference values where obtained by projection on the three spaces determined (one per block), with 8 ICs extracted from the 2008 spectral database (281.202 individuals). The results obtained shows correlations between SSC and the 6th -or 7th IC ranging from 0.61 to 0.67. The plot of these signals for the three blocks is shown in Figure7 and they show maxima at 980 nm and 1180 nm. These regions were identified as SSC informative by the original MLR model.

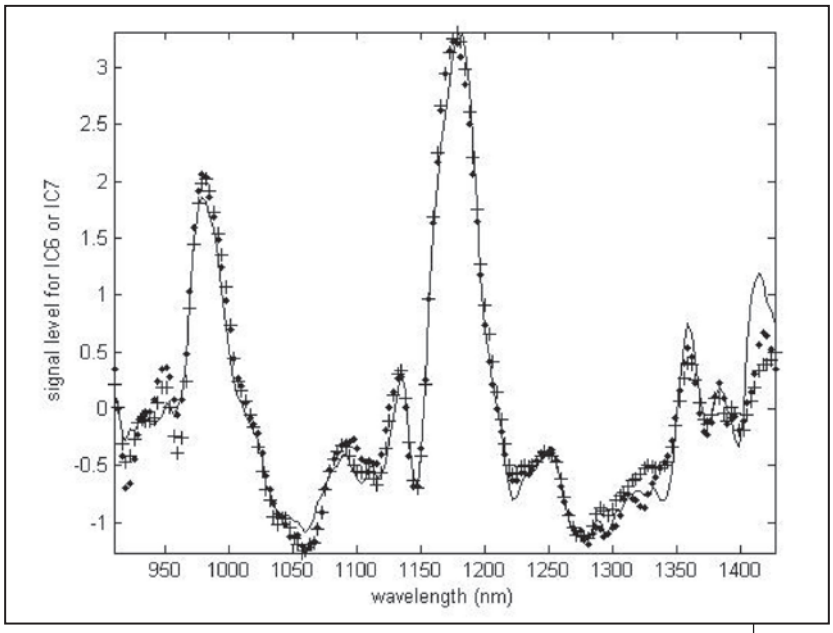

Figure 7: *Signals extracted from the three different blocs. Block 1 - IC7 (solid line), block 2 –IC6 (pluses) and block 3 – IC6 (dotted line). 8 ICs extracted from the 2008 dataset (281.202 samples)*

CONCLUSION

The number of true ICs revealed by the *ICA-By-Blocks* procedure for each season separately fits on what could be expected from the practical knowledge of the system. The fact that only two significant ICs were determined when the complete dataset was analyzed may imply that sources of variation are changing through the seasons in a way that they were identified as different ICs.

High correlation between SSC and the scores obtained for the 6th or 7th IC extracted for the different blocks were found for 2008 season. This fact points to the possibility of improving the classification by the use of some ICs instead of the original spectra.

Further analysis has to be done for the possible identification of the extracted ICs, for the complete dataset and the data of each season, with the experience based sources of variation, and especially with SSC.

References

[1] P. Barreiro, A. Moya-González. Calibration transfer techniques and spectra control in the framework of breeding processes, Afrodata 2010, Rabat (Morocco), 2010.

[2] P. Barreiro, F. Chauchard, J. M. Roger, A Moya-Gonzalez, and V. Bellon-Maurel. Robust modelling for at-line and on-line calibration transfer in a NIR industrial application. Chimiometrie. Lille (France), 2005.

[3] A. Moya-González, P. Barreiro, J.M. Roger, B. Diezma, and J. Ortiz-Cañavate. Procedure for calibration transfer between seasons for on-line NIR evaluation of SSC in onion breeding lines. International Conference on Agricultural Engineering, Clermont-Ferrand (France), 2010.

[4] J.F. Cardoso, A. Souloumiac, «Blind beamforming for non-Gaussian Signals», IEE proceedings-F 140 (6) (1993), 362-370